

READ INSTRUCTIONS REPORT DOCUMENTATION PAGE BEFORE COMPLETING FORM EPORT NUMBER 2. GOVT ACCESSION NO. 3. RECIPIENT'S CATALOG NUMBER

A-3

A-3

ITLE (and Substitle)

SELF-CRITICAL, AND ROBUST, EST PARAMETERS OF THE MULTIVARIATE BUTION

ON TOTAL CONTROL OF THE MULTIVARIATE BUTION

ON TOTAL CONTR SELF-CRITICAL, AND ROBUST, ESTIMATES FOR THE PARAMETERS OF THE MULTIVARIATE NORMAL DISTRI-

5. TYPE OF REPORT & PERIOD COVERED Interim Technical Report

6. PERFORMING ORG. REPORT NUMBER A-3

B. CONTRACT OR GRANT NUMBER(a)

DAA G29-81-K-0110

Rensselaer Polytechnic Institute Troy, New York 12181

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

CONTROLLING OFFICE NAME AND ADDRESS

Approved for public release; distribution unlimited.

12. REPORT DATE 1 June 1982

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS(If different from Controlling Office)

Department of the Navy Office of Naval Research 715 Broadway (5th Floor) New York, New York 10003 15. SECURITY CLASS. (of this report)

15a. DECLASSIFICATION/DOWNGRADING

16. DISTRIBUTION STATEMENT (of this Report)

U. S. Army Research Office Post Office Box 12211 Research Triangle Park. NC 27709

17. DISTRIBUTION STATEMENT (of the obstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS PEPORT ARE THOSE OF THE AUTHOR(S) AND SHOWED AS AN OFFICIAL DEPARTMENT OF THE ARMY FUS ED & FOLLOW OR DE-CISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

robust estimation, multivariate normal distribution, M-estimators, outlier identification

ABSTRACT (Continue an reverse side if necessary and identify by block number)

This algorithm yields joint robust estimates of the location vector and the variance-covariance matrix for samples from the multivariate normal distribution. The degree of robustness depends on a single filtering parameter, c, set by the user. The algorithm provides, for each observation, an internally determined weight which may be used to identify potential outliers.

DD 1 1473 1473 EDITION OF ! NOV 65 IS OBSOLETE

# SELF-CRITICAL, AND ROBUST, ESTIMATES FOR THE PARAMETERS OF THE MULTIVARIATE NORMAL DISTRIBUTION

by

N. J. Delaney Union College

and

A. S. Paulson\*
Rensselaer Polytechnic Institute

<sup>\*</sup>Research supported in part by U.S. Army Research Office under contract DAA G29-81-K-0110.

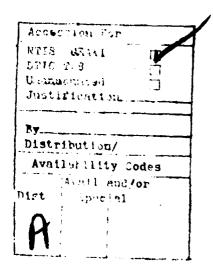
# Self-Critical, and Robust, Estimates for the Parameters of the Multivariate Normal Distribution

Keywords: Robust estimation; multivariate normal distribution;
M-estimators; outlier identification.

Language: ISO Fortran

Purpose: This algorithm yields joint robust estimates of the location vector and the variance-covariance matrix for samples from the multivariate normal distribution. The degree of robust-ness depends on a single filtering parameter, c, set by the user. The algorithm provides, for each observation, an internally determined weight which may be used to identify potential outliers.





### Theory:

The problems inherent in multivariate parameter estimation and the identification of potential outliers have been documented in Gnanadesikan (1977, Chapter 5) and Barnett and Lewis (1979, Chapter 6). Procedures for the simultaneous estimation of the location vector  $\mu$  and the covariance matrix V of the p-variate Gaussian distribution

$$f(x|\mu, v) = \frac{|v|^{-\frac{1}{2}}}{(2\pi)^{p/2}} \exp\{-\frac{1}{2}(x-\mu)^{T}v^{-1}(x-\mu)\},$$
 (1)

given a random sample  $x_1$ ,  $x_2$ , ...,  $x_n$ , have been discussed by Maronna (1976) and Huber (1977, Chapter 5). Our procedure stems from a single underlying primitive principle which reduces to likelihood as a special case and has desirable properties not possessed by competitors. Let c be a real number. The self-critical estimators  $\overline{\mu}$  and  $\overline{V}$  are determined from the zeros of

$$\sum_{i=1}^{n} \frac{f_{i}^{c}}{Q} \left\{ (1+c) \frac{\partial \log f_{i}}{\partial \mu} - \frac{1}{Q} \frac{\partial Q}{\partial \mu} \right\} = 0 , \qquad (2)$$

and

$$\sum_{i=1}^{n} \frac{f_{i}^{c}}{Q} \left\{ (1+c) \frac{\partial \log f_{i}}{\partial V} - \frac{1}{Q} \frac{\partial Q}{\partial V} \right\} = 0 , \qquad (3)$$

where

$$Q(\mu, V; c) = \int_{R_{p}} f^{1+c}(x|\mu, V) dx$$

$$= \{ (1+c)^{p} (2\pi)^{cp} |V|^{c} \}^{-\frac{1}{2}}.$$
(4)

The arguments of f and Q have been suppressed in (2) and (3) for notational convenience. Equations (2) and (3) may also be obtained from the maximization of

$$L_{c} = \frac{1}{c} \sum_{i=1}^{n} \left\{ \frac{f_{i}^{c}}{o^{c/(1+c)}} - 1 \right\}$$
 (5)

with respect to  $\overset{\upmu}{\overset{\upmu}{}}$  and  $\overset{\upmu}{\overset{\upmu}{}}$  (Paulson and Delaney, 1982). It may be shown that

$$\lim_{c \to 0} L_c = \sum_{i=1}^{n} \log f_i \tag{6}$$

so that the self-critical procedure reduces to maximum likelihood as a special case. These procedures may also be developed from consideration of the generalized mean. We have termed the procedure self-critical because the information component supplied by the expressions in brackets  $\{\cdot\}$  in (2) and (3) are "fed back" through the assumed density f with degree of criticism determined by c. Observations which receive relatively low final weights  $f_{1}^{C}/Q$ , a scale-free expression, are candidates for special examination as potential outliers.

Equations (2) and (3) reduce, on simplification, to the joint iterative forms

$$\overline{\mu}_{m+1} = \sum_{i=1}^{n} w_{mi} x_{i}, \qquad (7)$$

and

$$\overline{V}_{m+1} = (1+c) \sum_{i=1}^{n} w_{mi} (\overline{x}_{i} - \overline{\overline{u}}_{m}) (\overline{x}_{i} - \overline{\overline{u}}_{m})^{T}, \qquad (8)$$

where

$$\mathbf{w_{mi}} = \frac{\exp\{-\frac{c}{2} (\mathbf{x_{i}} - \overline{\mu}_{m})^{T} \overline{\mathbf{v}_{m}}^{-1} (\mathbf{x_{i}} - \overline{\mu}_{m})\}}{\sum_{i=1}^{n} \exp\{-\frac{c}{2} (\mathbf{x_{i}} - \overline{\mu}_{m})^{T} \overline{\mathbf{v}_{m}}^{-1} (\mathbf{x_{i}} - \overline{\mu}_{m})\}}$$
(9)

for  $m = 0, 1, ..., m^*$ . Initial estimates  $\overline{\mu}_0$  and  $\overline{v}_0$  are required to set the

iterative procedure embodied in (7), (8), and (9) in motion. We have frequently used  $\mathbf{w}_{m*i}$  instead of the final  $\mathbf{f}_{i}^{C}/Q$  to assess patterns in the data with respect to the multivariate Gaussian assumption and the internal consistency of the data.

It is recommended that the user experiment with the value c in examining a given data set. The estimators  $\overline{\mu}$  and  $\overline{V}$  are increasingly robust with increasing c because of the increasingly self-critical nature of the procedure. If for several values of c, similar estimates of  $\mu$  and V are obtained and no weight  $\mathbf{w}_{\mathbf{m}^*\mathbf{i}}$  is relatively small, then the data and Gaussianity are self-consistent. If the estimates of  $\mu$  and V vary substantially with c or at least one  $V_{\mathbf{i}}$  lue of  $\mathbf{w}_{\mathbf{m}^*\mathbf{i}}$  is small relative to the remainder, then those observations so labeled should be set aside for special examination vis- $\mathbf{a}$ -vis the remainder. Since the  $\mathbf{w}_{\mathbf{m}^*\mathbf{i}}$  induce a virtually automatic ranking of the observations, they are very useful in determining patterns in the data. We have typically used  $0 \le c \le 1$  for the multivariate Gaussian distribution although Paulson, Presser, and Nicklin (1982) have found use for negative values of c, as well as values of c in excess of unity. The magnitude to which c may be set is to a large extent dependent on the sample size.

The estimators  $\mu$  and  $\overline{V}$ , for all c>0, are location and scale invariant, are M-estimators, are jointly asymptotic normal, have closed, bounded and redescendent to zero influence functions, and are, for moderate values of c, relatively efficient (Paulson and Delaney, 1982).

# Numerical Method:

The algorithm begins by computing initial estimates  $\bar{\mu}_0 = \hat{\mu}$ , the usual vector of sample means, and  $\bar{V}_0 = \hat{V}$ , the maximum likelihood estimator of the dispersion matrix. Formulae (7), (8), and (9) are the basis for the self-critical iterative procedure. Successive estimates of  $\mu$  and V are generated until the norms of consecutive estimates are less than a user specified small amount, ETA, for three successive iterations. Generally, we have used ETA =  $10^{-4}$  or  $10^{-5}$ . Of course, the number of iterations required for convergence will depend on the value of ETA specified.

# Structure:

# SUBROUTINE SCEST(X,M,N,MDIM,NDIM,C,ETA,IMAX,MU,V,WT,MUZ,VZ,IT,IFAULT)

# Formal parameters

x	Real array	(M,N)	input:	data, M dimensions, N points
M	Integer		input:	the number of dimensions
N	Integer		input:	the number of sample points
MDIM	Integer		input:	the row dimension of X in calling program
NDIM	Integer		input:	the column dimension of X in calling program
С	Real		input:	filtering parameter
ETA	Real		input:	convergence criterion for norms of consecutive estimates
IMAX	Real		input:	maximum number of iterations desired
MU	Real array	(M)	output:	final location vector estimate, M dimensions
v	Real array	(M,M)	output:	final var-cov matrix estimate
WT	Real array	(N)	output:	final weight assigned to each sample point in computation of estimates
MUZ	Real array	(M)	output:	initial location estimate; vector of means
vz	Real array	(M,M)	output:	initial var-cov estimate; MLE
IT	Integer		output:	the number of iterations used
IFAULT	Integer		output:	<pre>0 if no errors in computation 1 if estimate of var-cov matrix is not positive definite 2 if accuracy test failed on inverse calculation 3 if convergence criteria not met in max desired iterations</pre>

#### Restrictions:

The number of dimensions must not exceed 10, unless the dimensions of dummy arrays VINV, TMP, MUSAV, VSAV, and VNORM are redeclared (see subroutines WEIGHT and CKNRM). This procedure requires a matrix inversion subroutine. As written, subroutine WEIGHT calls on the IMSL (International Mathematical and Statistical Library) subroutine LINV2F which inverts a matrix with specifiable accuracy = IDGT. The use of alternate matrix inversion routines would require rewriting this portion of the subroutine WEIGHT.

#### Precision:

The DOUBLE PRECISION version is recommended due to the iterative nature of the algorithm and the matrix inverse calculation.

## Time:

Experience with an IBM3033 computer has shown that the procedure averaged 0.03 seconds for a bivariate sample of size 20 and 0.08 seconds for a trivariate sample of size 25. These computations were done using DOUBLE PRECISION and a convergence criterion, ETA =  $10^{-5}$ .

#### References:

- 1. Barnett, V., and Lewis, T. (1978). Outliers in Statistical Data.
  New York: Wiley.
- 2. Gnandesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. New York: Wiley.
- 3. Huber, P.J. (1977). Robust Statistical Procedures. Philadelphia: SIAM.
- 4. Maronna, R.A. (1976). Robust M-estimators of Multivariate Location and Scatter. Annals of Statistics, 4, pp. 51-67.
- 5. Paulson, A.S., and Delaney, N.J. (1982). Self-Critical Procedures for the Multivariate Gaussian Distribution. Submitted to <u>Technometrics</u>.
- Paulson, A.S., Presser, M.A., and Nicklin, E.H. (1982). Self-Critical, and Robust, Procedures for the Estimation of Parameters. Submitted for publication.

APPENDIX

FORTRAN PROGRAM AND SUBROUTINES

```
REAL*8 X(5,150),MU(5),V(5,5),RT(150),MUZ(5),VZ(5,5)
    REAL*8 ETA, C, CSAV (5), TMP (5), TMP2 (5)
    DIMENSION VP (30)
    MDIM=5
    NDIM=150

DIMPUSIONS <= 5
</p>
 DRIVER INPUT:
                 N
                     # SAMPLE PTS<= 150,
                                             3
                                     NC # VALUES OF C INPUT
               MAY ITERS ALLOWED,
         IMAX
                CONVERG. CRITERION FOR MORES
    READ (5,900) N, M, IMAX, NC, ETA
900 FORMAT (413, E10.5)
DRIVER INPUT: CSAV ARRAY OF C VALUES
    READ(5,901) (CSAV(I), I=1,NC)
901 FOPMAT (5F5.3)
DRIVER INPUT:
                 VF ARRAY FOR FORMAT OF DATA CAPDS PUT IN ()
    READ (5, 302) VF
902 FORMAT (30A4)
DRIVER INPUT:
                          DATA ONE SAMPLE POINT PEP CAPP
                  X(I,J)
    DO 10 J=1, N
    READ(5, VF)(X(I, J), I=1, M)
 10 CONTINUE
    DO 100 KC=1,NC
    C=CSAV (KC)
    CALL SCEST (X,M,N,MDIM,NDIM,C,ETA,IMAX,MU,V,WT,MUZ,VZ,IT,IFAULT)
    WRITE (6,910)
IF (IFAULT. EQ. 9) GO TO 30
NONNORMAL TERMINATION MESSAGES
    GO TO (11,12,13), IFAULT
 11 WRITE (6,911) IT
911 FORMAT (1HO, 46HESTIMATE OF VAR-COV MATPIX ALGORITHMICALLY NOT
   & 19H POSITIVE DEFINITE, 4Y, 10HITER NO. =, 15)
    GO TO 30
 12 WRITE (6,912) IT
912 FORMAT (1HO, 45HACCURACY TEST PAILED ON INVERSE CALCULATION
   \varepsilon 10 HITER NO. =, 15)
    GO TO 30
 13 WRITE (6,913) IMAX
913 FORMAT (1HO, 32HCCNVERGENCE CRITERIA NOT MET IN ,15,2X,10HITERATIONS)
NORMAL TERMINATION AND OUTPUT
 30 WRITE (6, 930) C, ETA
930 FORMAT (1HO, 10X, 34H*** SELF-CRITICAL
                                            ESTIMATES
   624HFILTERING PARAMETER, C = , F7.4, 3X,
   \epsilon20HCCNVERG. CRITERION =, \epsilon12.5)
COMPUTE INITIAL + FINAL CORRELATION ESTIMATES
STORE IN LOWER OFF-DIAG OF VAR-COV FOR OUTPUT
    DO 35 I=1,M
    TMP2(I) = DSQRT(VZ(I,I))
    IF (IFAULT. EQ. 0) TMP (I) = DSQRT (V (I, I))
 35 CONTINUE
    K=M-1
    DO 45 I=1,K
    L=I+1
    DO 40 J=L, M
    V7.(J,I) = V2.(I,J) / (TMP2.(I) *TMP2.(J))
    IP (IFAULT. EQ. 0) V(J, I) = V(T, J) / (T^{MP}(J) *T^{MP}(J))
 40 CONTINUE
 45 CONTINUE
 PRINT OUT INITIAL ESTIMATES
```

WRITE (6,931) 931 FORMAT (1HO, 10x, 17HINITIAL ESTIMATES) DO 50 I=1.4 WRITE (6,932) MUZ (I), (VZ (I,J), J=1, M) 932 PORMAT (1H0, P10.5, 8X, 5 (P10.5, 3X)) 50 CCMTINUE IF (IT. EQ. 1. AND. IFAULT. GT. 0) GO TO 100 PRINT OUT FINAL SC ESTIMATES WRITE (6,933) IT 933 PORMAT (1HO, 10X, 26HPINAL ESTIMATES ITER NO. =, 15) DO 60 I=1, M ERITE (6,932) MU (I), (V(I,J),J=1,M)60 CONTINUE PRINT OUT FINAL SC WEIGHTS POR EACH SAMPLE POINT WRITE (6,934) 924 PCPMAT (1HO, 3X, 3HNO., 3X, 6HWEIGHT, 10X, 12HSAMPLE POINT) DO 70 J=1, N WRITE (6,935) J, WT (J), (X (I,J), I=1, M) 935 FORMAT (1H0, I5, 3x, E12, 5, 3x, 5F10, 5) 70 CONTINUE 100 CONTINUE STOP

END

```
C
C
       SUBFOUTINE SCEST (X, M, N, MDIM, NDIM, C, PTA, IMAY, MU, V, WT, MUZ, VZ, IT, IFAULT)
C
C
   CALCULATES SELF-CRITICAL ESTIMATES
C
       REAL*8 X (MDIM, NDIM), MU (MDIM), V (MDIM, MDIM), WT (MDIM)
       REAL*8 MUZ (MDIM) . VZ (MDIM, MDIM) .ZERO.ONE
       DATA ZERO, CNE/0.000, 1.000/
       IFAULT=0
   OBTAIN INITIAL ESTIMATES- MLE
       CALL INITL (X, M, N, MUZ, VZ, MDIM, NDIM)
       DO 10 I=1, M
       MU(I) = MUZ(I)
       DO 5 J=1, M
    5 V(I,J) = VZ(I,J)
   16 CONTINUE
  ITERATIVE FORMATION OF NEW ESTIMATES
       DO 100 IT=1, IMAX
       CALL REIGHT (X,M,N,MDIM,NDIM,C,MU,V,WT,IFAULT)
       IF (IFAULT.NE.O) PETURN
   LCCATION ESTIMATE
       DC 20 I=1, M
       MU(I) = ZERO
       DO 15 J=1, N
   15 MU(I) = MU(I) + X(I,J) + WT(J)
   20 CONTINUE
   VAR-COV ESTIMATE
       DO 40 I=1, M
       DO 35 J=I,M
       V(I,J) = ZERO
      DC 30 K=1,N
   30 V(I,J) = V(I,J) + (X(I,K) - MU(I)) + (X(J,K) - MU(J)) + WT(K)
       V(I,J) = V(I,J) * (C+ONE)
       Ir(I.NE.J)V(J,I)=V(I,J)
   35 CONTINUE
   40 CONTINUE
C CHECK CONVERGENCE CRITERIA
      CALL CKNPM (M, MDIM, ETA, MU, V, IT, IFLAG)
      IF (IFLAG. EQ. 1) RETURN
  100 CONTINUE
  CONVERG. CRIT. NOT MET IN MAX ITERATIONS DESIRED
      IFAULT=3
       RETUSN
      END
C
C
      SUBROUTINE INITL (X, M, N, MUZ, VZ, MDIM, NDIM)
   FORMS INITIAL ESTIMATES- MLE
C
C
      REAL*8 X (MDIM, NDIM), MUZ (MDIM), VZ (MDIM, MDIM)
      REAL*8 DENOM, ZERO, ONE, SUM
      DATA ZERO, ONE/O. ODO, 1. ODO/
      DENCM=DPLOAT (N)
      DO 10 I=1, M
      MUZ (I) =ZERO
      DO 5 J=1, Y
```

```
5 \text{ MUZ}(I) = \text{MUZ}(I) + \text{X}(I,J)
       MUZ(I) = MUZ(I) / DENOM
    10 CONTINUE
       DO 25 I=1.5
       DO 20 J=I,M
       SUM=ZERO
       DO 15 K=1, N
    15 SUM=SUM+X(I,K) *X(J,K)
       VZ(I,J) = (SUM-DENOM*MUZ(I)*MUZ(J))/DENOM
       VZ(J,I) = VZ(I,J)
   20 CONTINUE
   25 CONTINUE
       RETURN
       END
C
C
       SUBFOUTINE WEIGHT (X, M, N, MDIM, NDIM, C, MU, V, RT, I PAULT)
C
C
   FORMS WEIGHTS FOR EACH SAMPLE POINT
C
       REAL+8 X (MDIM, NDIM), MU (MDIM), V (MDIM, MDIM), WT (NDIM)
       REAL*8 VINV (10, 10), TMP (10, 10), WKAREA (50)
       REAL*8 R.SUM.ZEEC.TWO
       DATA ZERO, TWO/0.0D0, 2.0D0/
       DO 10 I=1, M
       DO 5 J=1,5
    5 TMP (I,J) = V(I,J)
   10 CONTINUE
   OBTAIN INVERSE OF CURRENT VAR-COV MATRIX ESTIMATE
C
C
   LINV2F IS IMSL SUBROUTINE WHICH INVERTS MATRIX WITH
C
   SPECIFIAPIE ACCURACY=IDGT
       IDGT=5
       IER=0
      CALL LINV2F (TMP, M, 10, VINV, IDGT, WKAPEA, IER)
       IF (IER. EQ. 0) GO TO 20
   ERRORS IN MATRIX INVERSION
       IF (IPR. GE. 129) IPAULT=1
      IF (IER. EQ. 34) IFAULT=2
      RETURN
C
   END OF SECTION REQUIRED WITH IMSL SUBROUTINE
   20 SUM=ZERO
      DO 35 J=1,N
      R=ZERO
       DO 30 I=1,M
      DO 25 K=1, M
   25 R=R+(X(I,J)-MU(I))*VINV(I,K)*(X(K,J)-MU(K))
   30 CONTINUE
      R=R*C/TWO
      IF (R.LT.ZERO) IFAULT=1
      IF (IFAULT. EQ. 1) RETURN
      IP(R.LT.174.)WT(J) = DEXP(-P)
      IF (R. GE. 174.) WT (J) = ZERO
      SUM=SUM+WT (J)
   35 CONTINUE
      DO 40 J=1, N
   40 WT(J) = WT(J) /SUM
      RETURN
      END
```

```
C
C
      SUBROUTINE CKNRM (M, MDIM, BTA, MU, V, IT, IPLAG)
   CHECKS CONVERGENCE CRITERIA
C
C
      REAL+8 MU (MDIM), V(MDIM, MDIM), MUSAV (10), VSAV (10, 10), VNFM (10)
      REAL*8 MUNRM, ZERO
      DATA ZERO/0.000/, MCTMU/0/, MCTV/0/
      IFLAG=0
      IF(IT.EQ.1) GC TO 20
   LOCATION NORM = SUM ABSOLUTE DIFFERENCE OF
C
                      CONSEC. ESTIMATE ELEMENTS
C
C
   VAR-COV. NORM = MAX ROW SUM ABS. DIFFERENCE
                     CONSEC. ESTIMATE ELEMENTS
      MUNPM=ZERO
      DO 10 I=1,M
      MUNRM=MUNRM+DABS (MUSAV (I) -MU (I))
       VNRM(I) = ZERO
      DO 5 J=1, M
    5 VNRM(I) = VNRM(I) + CABS(VSAV(I,J) - V(I,J))
   10 CONTINUE
      MAX =1
      DO 15 I=1, M
       IF (VNRM (I) . GT. VNRM (MAX) ) MAX=I
   15 CONTINUE
       IF (VNFM (MAX) . LT. ETA) MCTV=MCTV+1
       IF (VNEM (MAX) . GE. ETA) MCTV=0
       IF (MUNRM.LT. ETA) MCTMU=MCTMU+1
       IF (MUNRM. GE. ETA) MCTMU=0
   CONVERGENCE CRITERIA MET ON 3 CONSEC. ESTIMATES
       IF (MCTV. GT. 2. AND. MCTMU. GT. 2) IFLAG=1
   SAVE CUFFENT ESTIMATES
C
   20 DO 30 I=1,M
       MUSAV(I) = MU(I)
       DO 25 J=1, M
   25 VSAV(I,J) = V(I,J)
   30 CONTINUE
       RETURN
       END
```